Detecting English Writing Styles For Non-native Speakers

Rami Al − Rf ou'
Department of Computer Science
Stony Brook University
NY 11794, USA
ralrfou@cs.stonybrook.edu

ABSTRACT

Analyzing writing styles of non-native speakers is a challenging task. In this paper, we analyze the comments written in the discussion pages of the English Wikipedia. Using learning algorithms, we are able to detect native speakers' writing style with an accuracy of 74%. Given the diversity of the English Wikipedia users and the large number of languages they speak, we measure the similarities among their native languages by comparing the influence they have on their English writing style. Our results show that languages known to have the same origin and development path have similar footprint on their speakers' English writing style. To enable further studies, the dataset we extracted from Wikipedia will be made available publicly.

KEYWORDS: Stylometric Analysis, Wikipedia, Author attribution, Classification, Clustering.

1 Introduction

Stylometric analysis has important applications that cover deception detection, authorship attribution and vandalism detection (Harpalani et al., 2011; Ott et al., 2011). Analyzing English writing styles for non-native speakers is a harder task due to the influence of their native spoken languages. Such influence introduces a bias in the orthographical and syntactic errors made by the author and the choice of vocabulary (Koppel et al., 2005a).

Previous work, in this regard, focused on smaller datasets like International Corpus of Learner English (ICLE) (Koppel et al., 2005a,b; Argamon et al., 2009). This choice limits the number of the native languages targeted and the set of topics covered. It also focuses on features that might only appear in the writing styles of students, e.g. choice of words (Tsur and Rappoport, 2007; Zheng et al., 2003; Gamon, 2004). Syntactic features as subject-verb disagreement, mismatch of noun-number pairs and wrong usage of determiners were studied by (Wong and Dras, 2009). (Wong and Dras, 2010, 2011) used more sophisticated syntactic features as parse trees to examine the frequency of some distinguishable grammar rules.

Different from previous work, we use different data source. We extract users' comments from the English Wikipedia talk pages [1]. Our dataset is more challenging to study for the following reasons: First, the comments tend to be shorter in length than the articles from ICLE, and they cover a diverse spectrum of topics. Second, Wikipedia users represent a wider range of fluency in English. Moreover, the style of the comments is colloquial which limits the choice of features used. Finally, the targeted languages are more. Our contributions in this paper are:

- Analyzing common mistakes and patterns of non-native speakers' writing styles.
- Studying the similarities among languages using the English writing styles of the non-native speakers.
- Publicly available dataset composed of English Wikipedia users' comments.

This paper is structured as follows: Section 2 discusses various aspects of Wikipedia's structure and content. Section 3 describes the methodologies used to construct the dataset and filter the noise. In section 4, we discuss the experiments conducted. Finally, we conclude and present possible avenues of future research.

2 Wikipedia

Wikipedia is the de facto source of knowledge for internet users. Recently, it is has been extensively used to help solving different information retrieval tasks, especially the ones that involve semantic aspects (Milne and Witten, 2008). Wikipedia can be used to help common NLP tools to perform better; the size of the data and the diversity of authors and topics play a key role. Moreover, the sustained growth of Wikipedia content can bring performance gains with no additional complexity costs.

With more than 90 thousand active users and 4.4 million articles (in its English version), Wikipedia spans large number of topics. Wikipedia pages are saved under a control revision system that keeps track of users' edits and comments. Such resource presents a higher quality of data that is not achievable by the other commonly used sources of text as news, blogs and scientific articles.

Wikipedia has a complex database structure to serve its users. Therefore, extracting data is not trivial. Our goal is to first identify the language skills of users and then collect their

---

[1]Most Wikipedia pages have corresponding discussion pages which are called talk pages.

contributions. To identify the language skills, Wikipedia has an information box called *Babel* that users can add voluntarily to their profile pages in order to state their skills in different languages. A user can identify her native language and her skills in non-native languages on a scale of 0-5.

The task of collecting the contributions of a specific user is a more complex procedure. The differences among Wikipedia page revisions has to be generated and linked back to the user table. The resources we have are not sufficient enough to process such huge amount of data[2]. Instead we noticed that Wikipedia pages have accompanying talk pages where users discuss different aspects of the articles. In those pages, the style guideline encourages the user to sign her comments with her own signature that links back to her user page. The style of writing of these talk pages are less formal and technical than the main pages of Wikipedia and has more colloquial features.

3 Experimental Setup

In English Wikipedia, we found that around 60 thousand users specified their language skills, 47% of whom are English native speakers. The total number of comments found in the processed talk pages is around 12 million. Only 2.4 million comments have users with identified language skills. Since almost half of the users contribute to the talk pages, the number of users who make at least one comment is around 30 thousand.

Since we have large number of comments and users, we have to filter the user base to increase the quality of the gathered data. The rules, specified for this filter, are as follows:

- Group the users by their native languages and only consider the users from the 20 most frequently used languages.
- English native speakers can pick more fine grained categories, e.g. UK English, US English, etc. Only speakers under the US English category are selected.
- Users who specified more than one native language are excluded to help avoid improbable scenarios where users claim to be native in many languages.

The dataset after filtering the users constitutes of 9857 users and 589228 comments. Comments were filtered according to the following criteria:

- Comments need to have at least 20 tokens.
- Proper nouns are replaced by their Part of Speech (POS) tags to avoid bias toward topics.
- Non-ASCII characters are replaced by a special character to avoid bias toward non-English character usage in the comments.
- The classifier has the same number of comments for each of its classes. The two baseline classifiers are: the most common label and the random classifier. Each of these will have an accuracy of `1/(number of classes)`.
- The dataset is split into 70% training set, 10% development set and 20% as a testing set.

4 Experiments

4.1 Features

Given a training dataset, the comments are grouped by classes. The following n-grams are constructed for each class:

---

[2]Recent efforts were made to generate the diffs `http://dumps.wikimedia.org/other/diffdb/`

Experiment	Logistic Regression	Linear SVM
Non-native	74.45%	74.53%
Frequent	50.27%	50.26%
Families	50.81%	50.53%

Table 1: Accuracy of classification using different learning algorithms.

- 1-4 grams over the comments' words.
- 1-4 grams over the comments' characters.
- 1-4 grams over the part of speech tags of comments' words.

For each class, we will construct $3 * 4$ n-gram models. For each comment, we will construct a feature vector of the similarity scores between the comment and each of the n-gram models. Therefore, if a problem has six classes, $6 * 3 * 4 = 72$ features will be generated for each comment.

For example, the similarity scores (Sim) calculated for a comment (C) against the words n-grams models $words\_model(n)$.

$$Sim(C, n) = \sum_{x \in grams(C,n)} \log_2(count(x, n))$$

$$count(x, n) = \begin{cases} words\_model(n, word), & x \in words\_model(n) \\ 1, & x \notin words\_model(n) \end{cases}$$

Other features also include the relative frequency of each of the stop words mentioned in the comment. The 125 stop words are extracted from the NLTK stop words corpus(Loper and Bird, 2002). Moreover, the average size of words, the size of the comments and the average number of sentences are also included.

4.2 Native vs Non-native Experiment

This experiment aims to detect the non-native speakers writing styles. The classifier should be able to distinguish between comments by native speakers and other non-native speakers. All users with native language other than English are placed into one category. The number of comments used is around 322K. Table 1 shows that the linear SVM classifier reaches 74.53% accuracy, given the features explained in section 4.1.

The most informative features are word trigrams, word unigrams, word bigrams, word 4-grams, character bigrams, POS tag 4-grams, ordered by their importance. Table 2a shows the most correlated grams with native and non-native speakers.

Analyzing table 2a, we can notice that some of the n-grams indicate common grammatical mistakes in non-native speakers' writing styles. For example, word unigrams show that non-native speakers tend to use "earth" instead of "Earth". Character bigrams show that separating the comma from the previous word by a space is a common usage of punctuation for non-native speakers. The usage of determiners is a problematic issue for non-native speakers. In the word 4-grams, we can see a common mistake in the use of "the" before a proper noun. The over-usage of "the" can be validated by looking at the character bigrams where "th" appears. Word trigrams

Feature	Speaker	Gram
Words Trigrams	Non Native	article on NNP *comma* you have
	Native	used in the / in the middle
Words Unigrams	Non Native	scene, To, describing, earth, referenced, am
Words Bigrams	Non	You have, we do
	Native	of people
	Native	years of, I see / if you
Words 4-grams	Non Native	but I think that / by the NNP of
	Native	(or at least / NNP on NNP NNP / if you don't
Characters Bigrams	Non Native	th, e *space*, *space comma*
	Native	l-, at, ll
PoS 4-gram	Non	NNS , DT NN
	Native	MD VB VBN ,
	Native	, PRP VBZ JJ VBP NN IN NN

Feature	Speaker	Gram
Words Bigrams	Russian	prove that
	German	you leave
	Spanish	to reveal
	Dutch	reliance on
Words Unigrams	Dutch	refuting
	Spanish	timelines
	German	tie
Words Trigrams	Russian	stick to what
	Spanish	find out the
	Dutch	end up in
Characters 4-grams	French	hee *space*
	French	ownr
	Dutch	c/es
Words 4-grams	French	NNP as far as
	German	the NNP who
	Spanish	,etc),
	EN-US	Does anyone know if
PoS 4-gram	Dutch	RB CD -RRB- IN
	EN-US	VBD , RB DT
	EN-US	CD VBD NNS IN

(a) Non-native speakers experiment   (b) Most frequent languages experiment

Table 2: Correlated grams and speakers. For each class and each informative features the z-scores of each n-gram are calculated. The n-grams with the highest z-scores are reported in the table.

show that native speakers use "the" correctly in "in the middle" where we expect non-native speakers to use "in middle". Moreover, from the character bigrams, non-native speakers use "at" less than the native speakers which might suggest that they incorrectly use other articles in place of "at". Character bigrams show a trend in spelling mistakes where non-native speakers type single "l" instead of "ll". Another mistake is the less frequent usage of apostrophes " ' "; this can be traced to more frequent usage of "am" in word unigrams for non-native speakers and the appearance of "don't" in the word 4-grams for native speakers.

The more fluent the speaker is in English, the closer her writing style to the native speaker style. To test our basic intuition, we take advantage of the specific language fluency levels that the user specified in the available Wikipedia Babel box. We designed two different variations of the previous experiment. In the first, we limited the non-native speaker's class to the speakers with basic English skills (identified by the fluency scale 0-2). Whereas the second variant is composed of the more advanced non-native speakers with English fluency levels ranging from scale 3-5. Figure 1 shows the classifier error rate in the previous three variations.

The increase in the error rate of classification confirms our intuition. Moreover, it increases our confidence in the information given by the users, regarding their language skills, in their profiles.
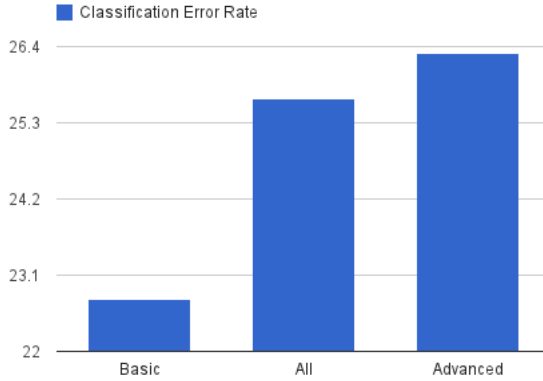
Figure 1: Classification error rate against non-native speakers with different skills.

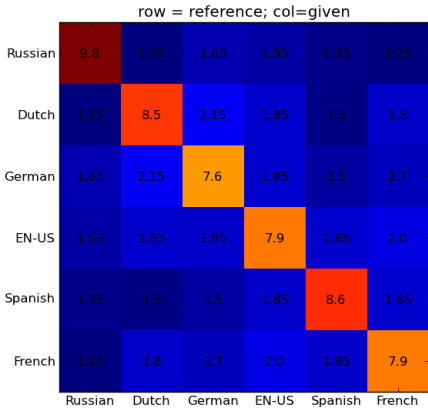## 4.3 Frequent Languages Experiment

This experiment aims to classify the comments written by the speakers of the most frequent native languages. Six languages are selected: US-EN, German, Spanish, French, Russian and Dutch. Figure 2a shows the confusion matrix of the logistic regression classifier. Table 1 shows that the best accuracy that the classifier achieved is 50.27% with 150K comments used. Looking at Figure 2a, we can see clearly that the Russian users are the easiest to identify. Moreover, the classification error is the highest in distinguishing the German and the Dutch users. These numbers confirm a basic intuition that the languages that have geographical proximity will have more borrowed words and grammars among them. Accordingly, this will affect their speakers' writing styles in English.

The most informative features are ordered as follows: word bigrams, word unigrams, word trigrams, character 4-grams, word 4-grams, POS tag 4-grams. We can see that the features of the longer grams become less informative, once we increased the number of classes given to the classifier because of sparsity. It may also indicate that there is an influence of the comment's topic on the classification. Table 4.3 shows some different mistakes made by different native speakers. For example, for French speakers "ownr" was a common mistake and not adding the space after the comma was another one among the Spanish speakers.
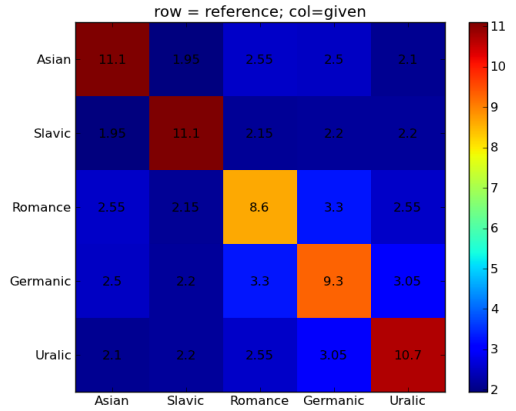
4.4 Languages Families Experiment

Looking at experiment 4.2, the confusion in classifying Dutch and German users suggests that there is a similarity between groups of languages. Referring to the linguistics research history of classifying the languages into families according to similar features and development history, this experiment validates such grouping. The following 18 languages are grouped into 5 families as:

- **Germanic**: German, Dutch, Norwegian, Swedish, Danish.
- **Romanace**: Spanish, French, Portuguese, Italian.
- **Uralic**: Finnish, Hungarian.

(a) Frequent languages experiment   (b) Languages families

Figure 2: Confusion Matrices of different Experiments

- **Asian**: Mandarin, Cantonese, Japanese, Korean.
- **Slavic**: Russian, Polish

Figure 2b shows that the Slavic and Asian native speakers have a clear English writing style which is easier to detect. The highest confusion in classification is between the Germanic and Romance languages, where geographical proximity plays a role in similarity. With the same reasoning, we can see the confusion between Germanic and Uralic languages.

Taking the opposite approach, we took the speakers of the most frequent 20 native languages and applied the same classification procedure over the new classes. The accuracy of the classifier is 25%. However, considering the confusion matrix as a similarity matrix, we applied the affinity propagation clustering algorithm (scikit learn developers, 2011) over the confusion matrix and the clusters that were formed are the following:

- **Cluster 1**: Arabic.
- **Cluster 2**: Danish, Dutch, Finnish, Norwegian, Swedish.
- **Cluster 3**: French, Italian, Portuguese, Spanish.
- **Cluster 4**: Mandarin, Cantonese, Japanese, Korean.
- **Cluster 5**: Russian, Polish, Turkish.
- **Cluster 6**: Hungarian, German, US-EN.

The above clusters, to large extent, support the literature classification of languages. Scrutinizing the 4-grams POS tags reveal more interesting observations regarding non-native speaker's usage of English. For example, in Portuguese speakers' comments, the pattern IN DT NN PRP appears 0.13% of the total number of their POS 4-grams. However, it only appears 0.04% in the Korean speakers' comments. Another observation can be seen by looking at the NN NN IN DT's usage in Portuguese and Polish comments. It appears 0.15% in the former POS 4-grams but only 0.05% in the later. Arabic speakers tend to use the pattern TO DT NN IN so frequently that it appears 0.11% while it is less than 0.06% for other speakers' POS 4-grams distributions. Finally, Japanese and Danish speakers slightly prefer the pattern NN PRP VBZ RB more than others.
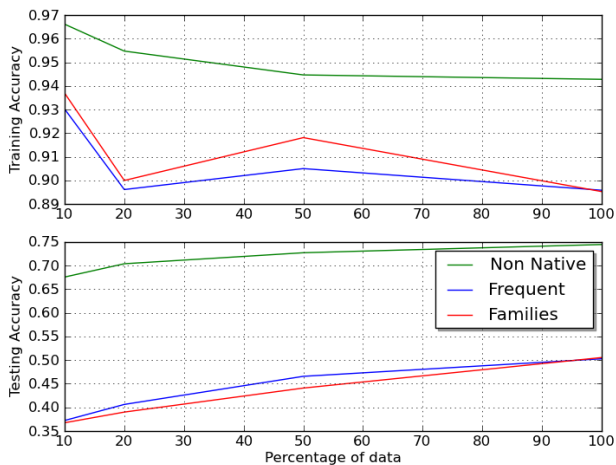
Figure 3: Learning curves of the logistic regression algorithm.

4.5 Learning Algorithms

Figure 3 shows a typical over-fitting situation where more data you have, the better rate the classifier can achieve. Here, the size of data that can be extracted from Wikipedia plays a significant role to boost the accuracy from 37% to over 50% in case of the Frequent languages and Families of languages experiments. This confirms the importance of the coverage of words in the language models that supply the frequency counts on the performance of the classifier.

Conclusions and Future Work

Our results show the effectiveness of the features constructed in detecting writing styles of a challenging and diverse content. The robustness of the features helped us in building competitive classifiers on different tasks. Moreover, we were able to analyze common users' usage patterns of English and discover grammatical and spelling errors.

Different languages showed different effects on the writing styles of their speakers. We were able to identify such trends between users' writing styles and cluster them into groups that supported the well studied origins of languages.

The learning curves show that it is worth increasing the size of the data in order of magnitude by adding the Wikipedia diffs, especially the non-minor ones, as it represents another source of users' contributions.

Acknowledgements

References

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Harpalani, M., Hart, M., Signh, S., Johnson, R., and Choi, Y. (2011). Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. In *ACL (Short Papers)*, pages 83–88. The Association for Computer Linguistics.

Koppel, M., Schler, J., and Zigdon, K. (2005a). Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*, pages 41–76.

Koppel, M., Schler, J., and Zigdon, K. (2005b). Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628. ACM.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *CoRR*, cs.CL/0205028.

Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

scikit learn developers (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

Tsur, O. and Rappoport, A. (2007). Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16. Association for Computational Linguistics.

Wong, S.-M. J. and Dras, M. (2009). Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*, page 53.

Wong, S.-M. J. and Dras, M. (2010). Parser features for sentence grammaticality classification. In *Proc. Australasian Language Technology Association Workshop*.

Wong, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Zheng, R., Qin, Y., Huang, Z., and Chen, H. (2003). Authorship analysis in cybercrime investigation. *Intelligence and Security Informatics*, pages 959–959.